12. Clancey, W.J. The epistemology of a rule-based expert system. Memo HPP-81-17, Heuristic Programming Project, Stanford University, December 1981. To appear in Artificial Intelligence, 1983.

13. Shortliffe, E.H. Clinical consultation systems: designing for the physician as computer user. Proceedings of the Fifth Symposium on Computer Applications in Medical Care, pp 235-6, Washington, D.C., November 1981.

14. Shortliffe, E.H. Problems in implementing the computer for continuing education. Presented at Conference on the Practical Impact of the Computer and Other Electronic Technology on the Future of Continuing Education, Palm Springs, California, January 1982. To appear in Mobius, 1982.

15. Wallis, J.W. and Shortliffe, E.H. Explanatory power for medical expert systems: studies in the representation of causal relationships for clinical consultations. Submitted for publication, December 1981.

16. Shortliffe, E.H. Technology and the hospital ward. To appear in Implementing Medical Technology: Issues and Strategies (S.R. Reiser and M. Anbar, eds.), 1982.

17. Shortliffe, E.H. and Fagan, L.M. Expert systems research: modeling the medical decision making process. Proceedings of the Workshop on Integrated Approaches to Patient Monitoring, University of Florida, Gainesville, Florida, March 1982. Also available as Memo HPP-82-3, Stanford Heuristic Programming Project, Stanford University, Stanford, California 94305.

18. Shortliffe, E.H. Computer-based clinical decision aids: some practical considerations. Proceedings of the AMIA Congress 82, San Francisco, CA., May 2-5, 1982.

19. Shortliffe, E.H. The Computer and Medical Decision Making: Good Advice is not Enough. Guest Editorial to appear in IEEE Engineering in Medicine and Biology Magazine, June 1982.

20. Gerring, P.E. The Interviewer/Reasoner Model: An architecture for interactive AI systems. Submitted to Proceedings of AAAI-82, Pittsburgh, Pennsylvania, 1982.

21. Clancey, W.J. Methodology for Building an Intelligent Tutoring System. To appear in Problems of Methods and Tactics in Cognitive Science. Kintsch, Polson, and Miller, Eds., 1982.

22. London, B. & Clancey W. J. Plan Recognition Strategies in Student Modeling: Prediction and Description. Submitted to AAAI-82.

E. Funding Support

Contract Title: "Exploration of Tutoring and Problem-Solving Strategies"
Principal Investigator: Bruce G. Buchanan
Agency: Office of Naval Research and
Advanced Research Projects Agency (joint)
ID number: N00014-79-C-0302
Term: March 1979 to March 1982
Total award: $396,325
[This project has recently been renewed for another three years. It will be jointly funded by the ONR and the ARI (Army Research Institute). The contract number has not yet been assigned, and the precise budget is still undergoing revision.]

Grant Title:"Research Program: Biomedical Knowledge Representation"
Principal Investigator: Edward A. Feigenbaum
Co-Principal Investigator (ONCOCIN Project): Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: LM-03395
Term: July 1979 to June 1984
Total award: $497,420

Grant Title: "Symbolic Computation Methods For Clinical Reasoning"
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: LM-00048
Term: July 1979 to June 1984
Total award: $196,425

II. INTERACTION WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations and Program Dissemination via SUMEX

A great deal of interest in MYCIN, GUIDON, ONCOCIN, and EMYCIN has been shown by the medical and computer science communities. We are frequently asked to demonstrate these programs to Stanford visitors or at meetings in this country or abroad. For example, in March 1982 Dr. Clancey demonstrated MYCIN in Helsinki, Finland for physicians at the University of Helsinki. Physicians have generally been enthusiastic about these programs' potential and what they reveal about current approaches to computer-based medical decision making. In all cases, the demonstrations have been performed on-line using network access to the SUMEX computer. The TYPER program, originally developed for the 1980 AIM Workshop by SUMEX staff in collaboration with Dr. Larry Fagan, has continued to be used to good effect for system demonstrations. As recently as May 1982 it was used when a group of physicians from the first meeting of the American Medical Informatics Association (held in San Francisco) visited Stanford to learn about SUMEX and our AIM research.

EMYCIN has generated considerable interest in the academic and business communities. We receive frequent requests for copies of the system and have been distributing it on a non-exclusive basis to anyone who wishes a copy.

NEOMYCIN was presented by Dr. Clancey at a Cognitive Science methodology conference sponsored by the Sloan Foundation in Boulder, Colorado in August 1981. Presentations of this kind carry SUMEX-AIM results out to cognitive psychologists from around the country. Dr. Clancey also presented GUIDON research to a seminar in the Stanford School of Education. Following directly from these contacts, Dr. Derek Sleeman, a computer scientist from Leeds, England with experience in computer-based education, will be arriving later in 1982 to coordinate collaboration between Education and Computer Science researchers at Stanford. We expect that several professors in Education, in particular Dr. Decker Walker, will become involved in our experiments with medical students and begin to participate in our study of medical education.

Several medical school and computer science teachers have also asked to use MYCIN in their computer science or medical computing courses, and we continue to make the programs available frequently to researchers around the world who access SUMEX using the GUEST account.

B. Sharing and Interaction with Other SUMEX-AIM Projects

We have continued collaboration with the PUFF and VM projects. Our development of a domain-independent system (EMYCIN) is facilitated by having a number of very different working systems on which to test our additions and modifications to the program. All the projects have provided us with useful comments and suggestions. Similarly, several researchers around the country, such as Professor Szolovits and students at MIT, have studied MYCIN in detail by running the system on SUMEX and interrupting to analyze the code and reasoning at key steps in a consultation. Thus MYCIN continues to be a vehicle for community building and a stimulus to the application of newer AI techniques that have been used to reimplement portions of MYCIN for experimental purposes.

The community created on the SUMEX resource has other benefits that go beyond actual shared computing. Because we are able to experiment with other developing systems, such as INTERNIST, and because we frequently interact with other workers (at the AIM Workshop or at other meetings around the country), many of us have found the scientific exchange and stimulation to be heightened. Several of us have visited workers at other sites, sometimes for extended periods, in order to pursue further issues which have arisen through SUMEX- or Workshop-based interactions. In this regard, the ability to exchange messages with other workers, both on SUMEX and at other sites, has been crucial to rapid and efficient exchange of ideas. Certainly it is unusual for a small community of researchers with similar scholarly interests to have at their disposal such powerful and efficient communication mechanisms, even among those on opposite coasts of the country.

C. Critique of Resource Management

The SUMEX facility has maintained the high standards that we have praised in the past. The staff members are always helpful and friendly, and work as hard to please the SUMEX community as to please themselves. As a result, the computer is as accessible and easy to use as they can make it. More importantly, it is a reliable and convenient research tool. We extend special thanks to Tom Rindfleisch for maintaining high professional standards for all aspects of the facility.

Finally, we continue to feel the need for more computing power. Much of our research and development continues to take place in the hours from 7 p.m. to 10 a.m., but it is unreasonable to expect all our programming staff to adjust their own schedules around a computer. The existence of the 2020 has been helpful in permitting demonstrations with good response time, and it has allowed us to introduce ONCOCIN in a real clinical environment, but ongoing R&D on the main machine remains difficult much of the time. Even the evening hours are now seeing higher load averages than was once the case. The addition of personal workstations is helping somewhat, but a great deal of research continues to require access to the central SUMEX machine. In this regard, we must note that the KI-10 hardware is becoming increasingly outmoded and is much less capable of managing the load placed upon it than would be the case with some of the newer large machines (e.g. a DEC 2060). Response time aside, we have shifted our development of GUIDON to the Xerox Dolphin in order to take advantage of the larger address space. This has also freed up disk space so that we can comfortably develop NEOMYCIN on SUMEX.

III. RESEARCH PLANS

A. Project Goals and Plans

EMYCIN

As we have noted, EMYCIN is now completed and we do not contemplate further work on the program. The research elements have largely shifted to the NEOMYCIN program under development by Dr. Clancey and colleagues as part of the GUIDON effort. During the coming year, therefore, we expect no direct work on EMYCIN but will be continuing two of the associated projects described in the progress report above:

(1) development of ROGET will continue and is likely to be complete by the end of 1982;

(2) the retrospective analysis of MYCIN/EMYCIN will be written by Buchanan and Shortliffe; the book should be largely complete and a publisher identified by this time next year.

GUIDON

Some of the long-range (2-3 years) goals of NEOMYCIN/GUIDON research are summarized below:

(1) to extend NEOMYCIN's model of diagnostic strategy to include common, non-expert approaches. Besides improving the program's ability to model the student, this enumeration of the space of strategies will allow us to follow a plan of research similar to Brown's and Burton's, but in the domain of diagnostic strategy as opposed to subtraction procedures. Eventually, we want to develop a principled psychological model that will relate strategies to knowledge and processing abilities.

(2) further studies of expert reasoning in domains that require "forming a picture" of a malfunctioning process. Experience with NEOMYCIN showed that expert diagnosticians attempt to order the data they collect causally, on a time line. Interpretation of observations can be partially understood as an attempt to match this description of onset, course, severity (intensity, frequency), and causal relations of findings onto known malfunctions that are recalled (indexed) by these process variables. This work will build upon recent advances in understanding causality (e.g., deKleer and Brown).

(3) exploitation of new technology for experimentation with teaching methods. How can we take advantage of the Dolphin's graphic capabilities in a GUIDON tutorial? Besides graphically presenting rule relationships, we might show the student the same kind of diagrams that we use when describing our knowledge bases to our AI colleagues (hierarchies, diagrams relating compiled associations to underlying causal chains). Other than presentation strategies, we would like to experiment with different interfaces, perhaps to break away from a continuous dialogue to use the screen more as a work space for annotating and examining the knowledge base, and organizing data and hypotheses in a diagnostic problem.

(4) incorporate GUIDON as an integral part of the curriculum in in medical diagnosis at Stanford. We propose to make GUIDON available at the Fleischmann Learning Center at Stanford Medical School, just as the traditional programs built at Massachusetts General and Ohio State were made available. In addition, we will work with one or more teaching fellows at the medical school to include GUIDON as part of the "physical diagnosis" course which is taught regularly at Stanford. This will continue our commitment to empirical research to develop our model of diagnosis and the teaching procedures.

ONCOCIN

There are five areas in which we expect to expend our efforts on the ONCOCIN System during the next few years:

(1) We will complete the three ONCOCIN evaluations described earlier.

(2) We will spend time improving the system's documentation and will prepare formal technical reports as well as clinical reports on the results of the evaluation studies.

(3) We will continue to develop the hypothesis assessment approach to consultation that was briefly described above.

(4) We will continue development of the query system and integration of the rule analysis system to aid in knowledge base development.

(5) We will begin encoding additional protocols if time permits.

We shall continue to relate the requirements of the system we are developing to the underlying artificial intelligence methodologies. We are convinced that the basic science frontiers of AI are best explored in the context of systems for real world use; thus ONCOCIN serves as a vehicle for developing an improved understanding of the issues that underlie all forms of knowledge engineering.

B. Requirements for Continued SUMEX Use

All the work we are doing (EMYCIN, GUIDON, ONCOCIN, plus continued use of the original MYCIN program) is totally dependent on continued use of the SUMEX resource. Although some of the GUIDON and ONCOCIN work is shifting to Xerox Dolphins, the SUMEX KI-10's and the 2020 continue to be key elements in our research plan. The programs all make assumptions regarding the computing environment in which they operate, and the ONCOCIN design in particular depends upon proximity to the DEC 2020 which enables us to use a 9600 baud interface.

In addition, we have long appreciated the benefits of GUEST and network access to the programs we are developing. SUMEX greatly enhances our ability to obtain feedback from interested physicians and computer scientists around the country. Network access has also permitted high quality formal demonstrations of our work both from around the United States and from sites abroad (e.g., Finland, Japan, Sweden, Switzerland).

We plan to continue development of NEOMYCIN and GUIDON on a SUMEX Dolphin that will be dedicated to this purpose. However, the project now includes 3 graduate students (one of whom is working jointly with Dr. Shortliffe on explanation research), a full-time programmer, and Dr. Clancey, so it will be necessary to continue use of the SUMEX DEC system for some of our programming (the development of NEOMYCIN's explanation system). Keeping our programs compatible with either system offers the advantage that they can be accessed over the network by remote users.

However, GUIDON2 will eventually be too large for the DEC system address space, so compatibility in the long run is only possible for NEOMYCIN.

C. Requirements for Additional Computing Resources

The acquisition of the DEC 2020 by SUMEX has been crucial to the growth of our research work, both to insure high quality demonstrations and to enable us to develop a system such as ONCOCIN for real-world use in a clinical setting. As we continue to develop systems that are potentially useful as stand-alone packages (e.g., an exportable EMYCIN), the addition of personal workstations has provided particularly valuable new resources. It is not yet clear which machines are optimal for the LISP-based applications we are developing, and an opportunity to test our systems on several small-to-medium machines will be invaluable and in keeping with our desire to move some of the AIM products into a community of service users.

As we have mentioned, the response time on the main machine continues to be a major problem, both during the daytime hours and frequently in the evenings as well. The SUMEX workstations have provided additional cycles to permit off-loading of some users from the PDP-10, and this has significantly benefited the SUMEX research community. In addition, we believe that our GUIDON and ONCOCIN experiences using the Dolphin personal computer are a significant part of our research. First, the Dolphin's large address space is permitting development of the large knowledge bases that these systems require. Second, the Dolphin's graphics will enable us to develop new methods for presenting material to naive users. Third, the Dolphin will provide a reliable, constant "load-average" machine, for running experiments with physicians and students. Finally, the development of ONCOCIN and GUIDON on the Dolphin will demonstrate the feasibility of running intelligent consultation or tutoring systems on small, affordable machines in physicians' offices, schools and other remote sites.

D. Recommendations for Future Community and Resource Development

Because the AI community has largely transitioned to DEC 20 systems running TOPS-20, it is becoming increasingly difficult to share software among sites. Even here at Stanford, there is considerable overhead in providing duplicate versions of our systems so that development work can be shared between SUMEX and the SCORE 2060. Utility software developed elsewhere also cannot be easily transferred to SUMEX for our use. In light of the age of the main SUMEX hardware and the divergence of our resource from related resources available elsewhere in the AIM and general AI communities, we would like to suggest consideration of an upgrade of SUMEX to a more modern central mainframe machine.

II.A.1.7 Protein Structure Project

Protein Structure Modeling Project

Prof. E. Feigenbaum and Mr. Allan J. Terry
Department of Computer Science
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Technical Goals

The goals of the protein structure modeling project are to 1) identify critical tasks in protein structure elucidation which may benefit by the application of AI problem-solving techniques, and 2) design and implement programs to perform those tasks. We have identified two principal areas which are of practical and theoretical interest to both protein crystallographers and computer scientists working in AI. The first is the problem of interpreting a three-dimensional electron density map. The second is the problem of determining a plausible structure in the absence of phase information normally inferred from experimental isomorphous replacement data. Current emphasis is on the implementation of a program for interpreting electron density maps (EDM's).

B. Medical Relevance and Collaboration

The biomedical relevance of protein crystallography has been well stated in an excellent textbook on the subject (Blundell & Johnson, Protein Crystallography, Academic Press, 1976):

> "Protein Crystallography is the application of the techniques of X-ray diffraction ... to crystals of one of the most important classes of biological molecules, the proteins. ... It is known that the diverse biological functions of these complex molecules are determined by and are dependent upon their three-dimensional structure and upon the ability of these structures to respond to other molecules by changes in shape. At the present time X-ray analysis of protein crystals forms the only method by which detailed structural information (in terms of the spatial coordinates of the atoms) may be obtained. The results of these analyses have provided firm structural evidence which, together with biochemical and chemical studies, immediately suggests proposals concerning the molecular basis of biological activity."

The project involves a collaboration between computer scientists at Stanford University and crystallographers at Oak Ridge National

Laboratories (Dr. Carroll Johnson), the University of California at San Francisco (Dr. Robert Langridge), and the University of California at San Diego (under the direction of Prof. Joseph Kraut). We also collaborate with Dr. Eric Grosse at Bell Laboratories, whose field is numerical analysis.

C. Progress Summary

We have spent most of the last year adding to the knowledge base. As a result of this work, CRYSALIS can now be considered a successful demonstration system. While the program is not mature enough to warrant its release to the general crystallographic community, it can solve some non-trivial proteins at a expert level of performance. Details can be found in publication nine.

CRYSALIS now consists of eight tasks. In the last year we have added two tasks for locating new islands of certainty in the hypothesis when the initial set has been used up. We have also implemented a task for computing a complete set of atomic coordinates from the set of superatoms found by CRYSALIS. In conjunction with this last task, we are developing FORTRAN algorithms for matching peptide and sidechain templates against the data to further refine the atomic locations.

Finally, we are compiling documentation on the system and the knowledge it embodies. These documents should be sufficiently complete so that we, or other groups, will have little difficulty picking up where we leave off. We also feel that explicit documentation of our model-building heuristics will be useful to the crystallographic community as it provides a new viewpoint, complementary to traditional crystallographic methods.

D. List of Publications

(1) Carroll Johnson and Eric Grosse, "Interpolation Polynomials, Minimal Spanning Trees, and Ridge-Line Analysis in Density Map Interpretation", American Crystallographic Association Program and Abstracts, 4:2, Evanston, Ill. Aug. 1976

(2) Robert S. Engelmore and H. Penny Nii, "A Knowledge-Based System for the Interpretation of Protein X-Ray Crystallographic Data," Heuristic Programming Project Memo HPP-77-2, January, 1977. (Alternate identification: STAN-CS-77-589)

(3) E.A. Feigenbaum, R.S. Engelmore, C.K. Johnson, "A Correlation Between Crystallographic Computing and Artificial Intelligence," in Acta Crystallographica, A33:13, (1977). (Alternate identification: HPP-77-15)

(4) Robert Engelmore and Allan Terry, "Structure and Function of the CRYSALIS System", Proc. 6IJCAI, 1979. pp250-256 (Alternative identification: HPP-79-16)

(5) R.S. Engelmore, A. Terry, S.T. Freer, and C.K. Johnson, "A Knowledge-Based System for Interpreting Protein Electron Density Maps", Abstracts of Amer. Crystallographic Ass. 7,1 (1979) p38

(6) E.H. Grosse, "Approximation and Optimization of Electron Density Maps", Stanford University Ph.D. Thesis, Dec. 1980 (Alternative identification: STAN-CS-80-835)

(7) R. Engelmore and A. Terry, Article VII.C3 (Crysalis) in Barr, A., and Feigenbaum, E. A. (eds.), The Handbook of Artificial Intelligence, Vol. II, Stanford Ca., HeurisTech Press, Los Altos, Ca.: Kaufman, 1982

(8) A. Terry and R. Engelmore, "A Knowledge-Based Approach to the Interpretation of Protein Electron Density Maps", in Machine Intelligence, Infotech State of the Art Report, Series 9, Number 3, Pergamon Infotech Ltd. Maidenhead, England, 1981

(9) A. Terry, "Hierarchical Control of Production Systems", Ph.D. Thesis, Dept. of Information and Computer Science, Univ. of Calif., Irvine, (forthcoming)

E. Funding status

Grant title: The Automation of Scientific Inference: Heuristic Computing Applied to Protein Crystallography

Principal Investigator: Prof. Edward A. Feigenbaum

Funding Agency: National Science Foundation

Grant identification number: MCS 81-17330

Term of award: January 15, 1982 through January 14, 1983

Amount of award: $28,976 (direct costs only)

II. INTERACTION WITH THE SUMEX-AIM RESOURCE

A. Collaborations

The protein structure modeling project has been a collaborative effort since its inception, involving co-workers at Stanford and UCSD (and, more recently, at Oak Ridge, UCSF, and Bell Laboratories). The SUMEX facility has provided a focus for the communication of knowledge, programs and data. Without the special facilities provided by SUMEX the research would be seriously impeded. Computer networking has been especially effective in facilitating the transfer of information. For example, the more traditional computational analyses of the UCSD crystallographic data are made at the CDC 7600 facility at Berkeley. As the processed data, specifically the EDM's and their Fourier transforms, become available, they are transferred to SUMEX via the FTP facility of the ARPA net, with a

minimum of fuss. (Unfortunately, other methods of data transfer are often necessary as well -- see below.) Programs developed at SUMEX, or transferred to SUMEX from other laboratories, are shared directly among the collaborators. Indeed, with some of the programs which have originated at UCSD and elsewhere, our off-campus collaborators frequently find it easier to use the SUMEX versions because of the interactive computing environment and ease of access. Advice, progress reports, new ideas, general information, etc. are communicated via the message and/or bulletin board facilities.

B. Interaction with Other SUMEX-AIM Projects

Our interactions with other SUMEX-AIM projects have been mostly in the form of personal contacts. We have strong ties to the MYCIN, AGE and MOLGEN projects and keep abreast of research in those areas on a regular basis through informal discussions. The SUMEX-AIM workshops provide an excellent opportunity to survey all the projects in the community. Common research themes, e.g. knowledge-based systems, as well as alternate problem-solving methodologies were particularly valuable to share.

C. Critique of Resource Services

The SUMEX facility provides a wide spectrum of computing services which are genuinely useful to our project -- message handling, file management, Interlisp, Fortran and text editors come immediately to mind. Moreover, the staff, particularly the operators, are to be commended for their willingness to help solve special problems (e.g., reading tapes) or providing extra service (e.g. immediate retrieval of an archived file). Such cooperative behavior is rare in computer centers.

There are several facilities we wish to single out as particularly useful in furthering our research goals. Since the members of the project are physically distant, the MSG program is very useful. Similarly, the file system, the ARCHIVE facility, and the general ease of getting backup files from the operator greatly aid our efforts at coordinating the efforts of collaborators using many large data sets and programs. The crystallographers in the project find SUMEX to be a friendly environment which allows them to do their work with a minimum of dealing with operating system details.

It has become increasingly evident, however, that as CRYSALIS expands, the facility cannot provide enough machine cycles during prime time to support the implementation and debugging of new features. For example, our segment-labeling preprocessor requires about an hour of machine time per 100 residues of protein (this is typically five to eight hours of terminal time during working hours) even when the LISP code is compiled.

III. USE OF SUMEX DURING THE REMAINING GRANT PERIOD (8/79 - 7/81)

A. Long-Range Goals

We have decided to end the project sometime in the fall. We have developed the system to where it demonstrates the basic soundness of our ideas. Unfortunately, development of the system into a generally-useful product would require the full-time assistance of an expert crystallographer. This person would need to restructure the knowledge base, pull the numerous FORTRAN data-reduction programs into a unified package, and extend the crystallographic knowledge contained in the system. We have not found a crystallographer willing to take over these responsibilities, so our goal is to thoroughly document the system and leave the code in a stable, runnable condition.

B. Justification for Continued Use of SUMEX

We feel that SUMEX is the ideal vehicle for further research on CRYSALIS. While some of our work is numerical in nature and uses such facilities as FORTRAN, our main interest is in artificial intelligence. Besides being an expert system of use to the crystallographic community, CRYSALIS is an exploration of the general signal processing problem. We are vitally concerned with issues such as proper architecture for using a wide variety of heuristics effectively and hypothesis formation when both data and model are poor. The utility of our work to the AI community is partially demonstrated by the development of the AGE project, an extension of Ms. Nii's early work on CRYSALIS.

This project progresses by the collaboration of several physically-separated groups. SUMEX provides a unique resource, an electronic community of researchers in our field, through the many systems such as net mail, country-wide access, and community workshops. We feel that CRYSALIS would not be possible outside of such a community.

While we plan to terminate active development work on CRYSALIS in the fall, we request continued use of SUMEX until the end of our grant. This period would be used for dissemination of results. We request use of SUMEX so that we can distribute programs to the crystallographic community and so that we can run demonstrations of our programs.

C. Needs and Plans for Other Computing Resources

We will make minor use of the Stanford Computer Science Department's SCORE machine, mostly for sending files to the Dover printer until such a facility is available on SUMEX.

D. Recommendations for Future Community and Resource Development

There are two recommendations we wish to make, the first and most important is to expand the computing power available to SUMEX users. CRYSALIS is an inherently-large problem. Proteins contain hundreds, to thousands of atoms which means large hypothesis structures, large

quantities of data, and a compute-bound inference program. As the system grows to maturity, we expect increasingly serious problems with address space limitations and with machine cycle availability.

The second recommendation is that SUMEX develop some relatively inexpensive file transfer facility for machines not on the ARPAnet. Software for this already exists in the form of the TTYFTP program (or possible future programs like it, but in a more portable language), the development needed is in hardware and in the TENEX operating system so that transfer rates greater than 1200 baud can be achieved. We are motivated to recommend this not only by our own need for such a facility, but also by the belief that it would aid other collaborations involving SUMEX and outside computers (the SECS project for example), and aid in the dissemination of useful programs from the research setting of SUMEX to user laboratories.

II.A.1.8 RX Project

The RX Project: Deriving Medical Knowledge from Time-Oriented Clinical Databases

Robert L. Blum, M.D., Ph.D.
Department of Computer Science
Stanford University

Gio C. M. Wiederhold, Ph.D.
Departments of Computer Science and Electrical Engineering
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Technical Goals

Introduction:

Medical and Computer Science Goals

The long range objectives of our project, called RX, are 1) to increase the validity of medical knowledge derived from large time-oriented databases containing routine, non-randomized clinical data, 2) to provide knowledgeable assistance to a research investigator in studying medical hypotheses on large databases, 3) to fully automate the process of hypothesis generation and exploratory confirmation. For system development we have used a subset of the ARAMIS database.

Computerized clinical databases and automated medical records systems have been under development throughout the world for at least a decade. Among the earliest of these endeavors was the ARAMIS Project, (American Rheumatism Association Medical Information System) under development since 1969 in the Stanford Department of Medicine. ARAMIS contains records of over 17,000 patients with a variety of rheumatologic diagnoses. Over 62,000 patient visits have been recorded, accounting for 50,000 patient-years of observation. The ARAMIS Project has now been generalized to include databases for many chronic diseases other than arthritis.

The fundamental objective of the ARAMIS Project as well as of all other clinical database researchers is to use the data that have been gathered by clinical observation in order to study the evolution and medical management of chronic diseases. Unfortunately, the process of reliably deriving knowledge has proven to be exceedingly difficult. Numerous problems arise stemming from the complexity of disease, therapy, and outcome definitions, from the complexity of causal relationships, from errors introduced by bias, and from frequently missing and outlying data. A major objective of the RX Project is to explore the utility of symbolic computational methods and knowledge-based techniques at solving some of these problems.

The RX computer program is designed to examine a time-oriented clinical database such as ARAMIS and to produce a set of (possibly) causal relationships. The algorithm exploits three properties of causal relationships: time precedence, correlation, and nonspuriousness. First, a Discovery Module uses lagged, nonparametric correlations to generate an ordered list of tentative relationships. Second, a Study Module uses a knowledge base (KB) of medicine and statistics to try to establish nonspuriousness by controlling for known confounders.

The principal innovations of RX are the Study Module and the KB. The Study Module takes a causal hypothesis obtained from the Discovery Module and produces a comprehensive study design, using knowledge from the KB. The study design is then executed by an on-line statistical package, and the results are automatically incorporated into the KB. Each new causal relationship is incorporated as a machine-readable record specifying its intensity, distribution across patients, functional form, clinical setting, validity, and evidence. In determining the confounders of a new hypothesis the Study Module uses previously "learned" causal relationships.

In creating a study design the Study Module follows accepted principles of epidemiological research. It determines study feasibility and study design: cross-sectional versus longitudinal. It uses the KB to determine the confounders of a given hypothesis, and it selects methods for controlling their influence: elimination of patient records, elimination of confounding time intervals, or statistical control. The Study Module then determines an appropriate statistical method, using knowledge stored as production rules. Most studies have used a longitudinal design involving a multiple regression model applied to individual patient records. Results across patients are combined using weights based on the precision of the estimated regression coefficient for each patient.

B. Medical Relevance and Collaboration

As a test bed for system development our focus of attention has been on the records of patients with systemic lupus erythematosus (SLE) contained in the Stanford portion of the ARAMIS Data Bank. SLE is a chronic rheumatologic disease with a broad spectrum of manifestations. Occasionally the disease can cause profound renal failure and lead to an early death. With many perplexing diagnostic and therapeutic dilemmas, it is a disease of considerable medical interest.

In the future we anticipate possible collaborations with other project users of the TOD System such as the National Stroke Data Bank, the Northern California Oncology Group, and the Stanford Divisions of Oncology and of Radiation Therapy.

We believe that this research project is broadly applicable to the entire gamut of chronic diseases that constitute the bulk of morbidity and mortality in the United States. Consider five major diagnostic categories responsible for approximately two thirds of the two million deaths per year in the United States: myocardial infarction, stroke, cancer, hypertension, and diabetes. Therapy for each of these diagnoses is fraught with controversy concerning the balance of benefits versus costs.

1) Myocardial Infarction: Indications for and efficacy of coronary artery bypass graft vs. medical management alone. Indications for long-term antiarrhythmics ... long-term anticoagulants. Benefits of cholesterol-lowering diets, exercise, etc.

2) Stroke: Efficacy of long-term anti-platelet agents, long-term anticoagulation. Indications for revascularization.

3) Cancer: Relative efficacy of radiation therapy, chemotherapy, surgical excision - singly or in combination. Optimal frequency of screening procedures. Prophylactic therapy.

4) Hypertension: Indications for therapy. Efficacy versus adverse effects of chronic antihypertensive drugs. Role of various diagnostic tests such as renal arteriography in work-up.

5) Diabetes: Influence of insulin administration on microvascular complications. Role of oral hypoglycemics.

Despite the expenditure of billions of dollars over recent years for randomized controlled trials (RCT's) designed to answer these and other questions, answers have been slow in coming. RCT's are expensive of funds and personnel. The therapeutic questions in clinical medicine are too numerous for each to be addressed by its own series of RCT's.

On the other hand, the data regularly gathered in patient records in the course of the normal performance of health care delivery are a rich and largely underutilized resource. The ease of accessibility and manipulation of these data afforded by computerized clinical databases holds out the possibility of a major new resource for acquiring knowledge on the evolution and therapy of chronic diseases.

The goal of the research that we are pursuing on SUMEX is to increase the reliability of knowledge derived from clinical data banks with the hope of providing a new tool for augmenting knowledge of diseases and therapies as a supplement to knowledge derived from formal prospective clinical trials. Furthermore, the incorporation of knowledge from both clinical data banks and other sources into a uniform knowledge base should increase the ease of access by individual clinicians to this knowledge and thereby facilitate both the practice of medicine as well as the investigation of human disease processes.

C. Highlights of Research Progress

1. 1 July 1981 to 1 May 1982

Our predominant objective was to detail the overall conceptual framework for the knowledge base and to develop the extensive computational machinery necessary for creating an adequate statistical study design, executing that study design on an on-line statistical package, and interpreting and incorporating the medical results.

The RX Knowledge Base (KB):

The central component of RX is a knowledge base of medicine and statistics, organized as a frame-based, taxonomic tree consisting of frames with attached data and procedures. Frames representing diseases and therapies contain procedures that use a variety of time-dependent predicates to label the patient records, facilitating the retrieval of time-intervals of interest in the records. Other frames representing statistical techniques are used to map hypotheses onto study designs and event definitions. Implementing the algorithms and data structures of this KB was one of the major tasks of the current year.

At the current time the RX KB contains about 250 frames of which 150 contain definitions and other relevant information pertaining to disease courses, effects of drugs, lab values, etc. This information compromises a small subset of medical knowledge dealing with some of the signs and symptoms of systemic lupus erythematosus (SLE) as well as the effects and indications of some drugs used for this disease. Other frames contain machine-readable knowledge of statistical techniques needed for testing entered hypotheses. There are approximately 40 time-dependent functions used to map from the database values onto defined frames.

The entire RX system currently contains approximately 400 INTERLISP functions accounting for 200 disk pages of code. The KB is about 60 disk pages. One disk page = 512 words * 36 bits per word. Also one disk page = approx. 1.5 typed pages on 8.5 by 11.5 inch paper.

Statistical Interfaces:

Once the relevant data have been abstracted from patient records they must be analyzed statistically. To do this we have recently adopted the IDL or Interactive Data-Analysis Language package developed at the Xerox Palo Alto Research Corp. IDL is a matrix manipulation language similar to APL and is built upon INTERLISP as is RX itself. The use of IDL for statistical analysis confers a tremendous advantage in that analyses are now highly interactive. IDL has completely supplanted our previous use of SPSS.

Time-Oriented Graphics Package:

This package enables data on an individual patient to be graphed over time, either linearly by visit or by calendar time with a "telescoping" capability. The program overlays graphs of both point data and data represented as episodes.

Clinical Study: The Effect of Prednisone on Cholesterol

As a testbed for the prototype system we have been investigating the hypothesis that the steroid, prednisone, produces a significant elevation of plasma cholesterol. To test this hypothesis, the records of 50 patients with systemic lupus erythematosus (SLE) were transferred from the ARAMIS Database to SUMEX. Of these patients, 18 were found to have five or more

cholesterol determinations and to have had sufficient variance in their prednisone regimens to be testable. The KB is used to elaborate a complex causal model for the prednisone/cholesterol hypothesis that is tested using a hierarchical multiple regression method with time-lagged values. The KB is used to determine sources of possible bias and to control for those variables in the regression or to eliminate corresponding time-intervals from records. An empirical Bayes method is used to average the estimated effects in patients with varying amounts of data.

The result, a highly statistically significant elevation of cholesterol by prednisone, will be submitted for publication during the coming year.

2. Research In Progress

Much work remains to be done in expanding the system software and in expanding the knowledge base. Current work is addressed to increasing the flexibility of the time-segmentation functions and enriching the data structures that encode relationships between objects.

We are trying to make increasingly general the class of medical hypotheses that the system can analyze automatically. This requires incorporating knowledge of additional statistical methods into the KB. We are also attempting to generalize our algorithms for selecting the set of variables that may potentially confound a given hypothesis. As a means for testing and expanding the system's capabilities we intend to perform several specific studies of importance in the management of the rheumatic diseases. Our study of the effect of prednisone on cholesterol was mentioned above. Other studies now being planned include the effect of chronic aspirin ingestion on liver function in rheumatoid arthritis, the specific incidence of infectious complications of steroids as a function of dose and duration, and the utility of various autoantibodies in the prediction of flares of SLE as compared to the utility of other indicators.

Finally, we are continuing to develop the Discovery Module in an attempt to make fuller use of the knowledge base to constrain the search space of hypotheses. This work is being pursued as doctoral dissertation research by Mark Erlbaum, a member of our group.

D. Publications

Blum, Robert L., Automated Induction of Causal Relationships from a Time-Oriented Clinical Database: The RX Project, Proceedings of the AMIA Congress, American Medical Informatics Association, San Francisco, 1982.

Blum, Robert L., Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Database: The RX Project, Lecture Notes in Medical Informatics, D.A.B. Lindberg and P.L. Reichertz (eds.), Springer-Verlag, in press. Blum, Robert L., Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Database: The RX Project, Computers and Biomedical Research, 15:2, 164-187, April, 1982.

Blum, Robert L., Discovery and Representation of Causal Relationships from a Large Time-Oriented Clinical Database: The RX Project, Doctoral Dissertation, Computer Science and Biostatistics, Stanford University, 1982.

Blum, Robert L., Displaying Clinical Data from a Time-Oriented Database, Computers in Biology and Medicine, 11:4, 197-210, 1981.

Blum, Robert L., Automating the Study of Clinical Hypotheses on a Time-Oriented Database: The RX Project, Proceedings of MEDINFO80, The Third World Congress of Medical Informatics, pp. 456-460, Tokyo, Oct. 1980 (also available as Stanford University Computer Science Dept. Report STAN-CS 79-816).

Blum, Robert L. and Wiederhold, Gio, Inferring Knowledge from Clinical Data Banks Utilizing Techniques from Artificial Intelligence, Proceedings of the Second Annual Symposium on Computer Applications in Medical Care, IEEE, Washington D.C., November, 1978.

Blum, Robert L., The RX Project: A Medical Consultation System Integrating Clinical Data Banking and Artificial Intelligence Methodologies, Stanford University Doctoral Dissertation Proposal, August 1978.

Wiederhold, Gio, Databases for Health Care, Lecture Notes in Medical Informatics, D.A.B. Lindberg and P.L. Reichertz (eds.), Springer-Verlag, 1981.

Wiederhold, Gio, Database Technology in Health Care, Journal of Medical Systems, 5:3, 175-196, 1981.

E. Funding Support Status

1) Integrating Medical Knowledge and Clinical Data Banks
Robert L. Blum, M.D.: Principal Investigator
National Library of Medicine, New Investigator Award
LM-03370
Total award: $90,000 (direct)
Term: July 1, 1979 through June 30, 1982

2) Deriving Knowledge from Clinical Databases
Gio C. M. Wiederhold, Ph.D.: Principal Investigator
National Center for Health Services Research
HS-4389
Total award: $127,000 (direct)
Term: September 30, 1981 through September 29, 1983

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Collaborations

Since our project is relatively new, we do not yet have public versions of the programs. There is, however, a large sphere of collaboration that we expect in the future. Once the RX program is developed, we would anticipate collaboration with some of the ARAMIS project sites in the further development of a knowledge base pertaining to the chronic arthritides. The ARAMIS Project at the Stanford Center for Information Technology is used by a number of institutions around the country via commercial leased lines to store and process their data. These institutions include the University of California School of Medicine, San Francisco and Los Angeles; The Phoenix Arthritis Center, Phoenix; The University of Cincinnati School of Medicine; The University of Pittsburgh School of Medicine; Kansas University; and The University of Saskatchewan. All of the rheumatologists at these sites have closely collaborated with the development of ARAMIS, and their interest in and use of the RX project is anticipated. We hasten to mention that we do not expect SUMEX to support the active use of RX as an on-going service to this extensive network of arthritis centers, but we would like to be able to allow the national centers to participate in the development of the arthritis knowledge base and to test that knowledge base on their own clinical data banks.

B. Interactions with Other SUMEX-AIM Projects

Several of the concepts incorporated into the design of the RX Project have been inspired by other SUMEX-AIM Projects. The RX knowledge base is similar to the Units Package of the MOLGEN PROJECT. The production rule inference mechanism used by us is similar to that in the MYCIN Project.

Several programs developed by the MYCIN group are regularly used by RX. These include disk hash file facilities, text editing facilities, and miscellaneous LISP functions. Regular communication on programming details is facilitated by the on-line mail system.

C. Critique of Resource Management

The SUMEX KI-10 has been severely overloaded for at least a year. Working in LISP is impossible during the day and is even difficult at times which were formerly low utilization times. This has forced us to rely increasingly on other local computation facilities.

The SUMEX resource management, per se, has always been accessible and cooperative in trying to provide our project with adequate resources subject to prevailing constraints.

III. RESEARCH PLANS

A. Project Goals and Plans

The overall goal of the RX Project is to develop a computerized medical information system capable of accurately extracting medical knowledge pertaining to the therapy and evolution of chronic diseases from a database consisting of a collection of stored patient records.

1. Short-Term Goals

Goals for the year August, 1982 through July, 1983 have been detailed in section IC. above on research in progress. To summarize that section, our main short-term goal is to generalize and refine our methods for representing causal relationships and for extracting those relationships from empirical data. We are further interested in refining our methods for constraining the hypothesis search space explored by our Discovery Module by more fully utilizing information from the knowledge base.

2. Long-Range Goals: August, 1982 through July, 1986

There are two inter-related long-range goals of the RX Project: 1) automatic discovery of knowledge in a large time-oriented database and 2) provision of assistance to a clinician who is interested in testing a specific hypothesis. These tasks overlap to the extent that some of the algorithms used for discovery are also used in the process of testing an hypothesis.

We hope to make these algorithms sufficiently robust that they will work over a broad range of hypotheses and over a broad spectrum of data distributions in the patient records.

B. Justification for Continued Use of SUMEX

Computerized clinical data banks possess great potential as tools for assessing the efficacy of new diagnostic and therapeutic modalities, for monitoring the quality of health care delivery, and for support of basic medical research. Because of this potential, many clinical data banks have recently been developed throughout the United States. However, once the initial problems of data acquisition, storage, and retrieval have been dealt with, there remains a set of complex problems inherent in the task of accurately inferring medical knowledge from a collection of observations in patient records. These problems concern the complexity of disease and outcome definitions, the complexity of time relationships, potential biases in compared subsets, and missing and outlying data. The major problem of medical data banking is in the reliable inference of medical knowledge from primary observational data.

We see in the RX Project a method of solution to this problem through the utilization of knowledge engineering techniques from artificial intelligence. The RX Project, in providing this solution, will provide an important conceptual and technologic link to a large community of medical

research groups involved in the treatment and study of the chronic arthritides throughout the United States and Canada, who are presently using the ARAMIS Data Bank through the CIT facility via TELENET.

Beyond the arthritis centers which we have mentioned in this report, the TOD (Time-Oriented Data Base) User Group involves a broad range of university and community medical institutions involved in the treatment of cancer, stroke, cardiovascular disease, nephrologic disease, and others. Through the RX Project, the opportunity will be provided to foster national collaborations with these research groups and to provide a major arena in which to demonstrate the utility of artificial intelligence to clinical medicine.

SUMEX as a Resource:

To discuss SUMEX as a resource for program development, one need only compare it to the environment provided by our other resource, the IBM 370/3081 installation at SCIP - the major computing resource at Stanford. Of the programs which we use daily on SUMEX -INTERLISP, MSG, TVEDIT, BBD, LINK - there is nothing even approaching equivalence on the 370, despite its huge user community. These programs greatly facilitate communication with other researchers in the SUMEX community, documentation of our programs, and the rapid interactive development of the programs themselves. The development of a program involving extensive symbolic processing and as large and complex as RX at the CIT facility, would require a staff many times as large as ours. The SUMEX environment greatly increases the productive potential of a research group such as ours to the point where a large project like RX becomes feasible.

Computation Resources Required by RX:

Disk Allocation:

RX requires the use of two large data files that need to be kept on-line: the patient database (DB) and the knowledge base (KB). In the course of testing a hypothesis several other files are used: inverted files, source files for statistical processing, LISP SYSOUT files, etc. Our current total disk allocation of 1500 pages for all RX group members has been just adequate. In the future, with anticipated expansions in numbers of patients and size of the KB, we intend to request an increase of our total allocation to 2000 pages.

C. Other Computational Resources

It is clear that the scope of potential application of the RX Project is large. Within the term of the SUMEX-AIM grant projected through July, 1986, we anticipate the involvement of several of the national ARAMIS collaborating institutions in developing and testing arthritis knowledge bases that reflect their own patient populations and therapeutic biases. The current SUMEX machine configuration will not be able to support this national interaction because the central processors of the KI-10 are already taxed to the limit. Ours is among the SUMEX groups that would

greatly benefit by the addition of one or more PDP-10 compatible machines, that could provide support to our anticipated national user community. Another resource that would be highly desirable is a faster and more reliable means for transferring data interactively between SUMEX and the CIT IBM 370. Our current method utilizes a 2400 baud line with transmission from CIT to SUMEX only, and is fraught with a high error rate. The addition of a reliable local network facility would greatly facilitate our ability to transfer patient files from CIT to SUMEX.

D. Recommendations for Resource Development

SUMEX is heavily loaded everyday and almost every evening. Program research is next to impossible during those periods. Program development would be greatly facilitated by the addition of any resources that lessened this loading: upgrading the current machine to a KL or 20/60 or adding core to decrease page swapping.

II.A.2 National AIM Projects

The following group of projects is formally approved for access to the AIM aliquot of the SUMEX-AIM resource or the Rutgers-AIM resource. Their access is based on review by the AIM Advisory Group and approval by the AIM Executive Committee.

II.A.2.1 Acquisition of Cognitive Procedures (ACT)

Acquisition of Cognitive Procedures (ACT)

Dr. John Anderson
Carnegie-Mellon University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

To develop a production system that will serve as an interpreter of the active portion of an associative network. To model a range of cognitive tasks including memory tasks, inferential reasoning, language processing, and problem solving. To develop an induction system capable of acquiring cognitive procedures with a special emphasis on language acquisition and problem-solving skills.

B. Medical Relevance and Collaboration

1. The ACT model is a general model of cognition. It provides a useful model of the development of and performance of the sorts of decision making that occur in medicine.

2. The ACT model also represents basic work in AI. It is in part an attempt to develop a self-organizing intelligent system. As such it is relevant to the goal of development of intelligent artificial aids in medicine.

We have been evolving a collaborative relationship with James Greeno and Allan Lesgold at the University of Pittsburgh. They are applying ACT to modeling the acquisition of reading and problem solving skills. We have made ACT a guest system within SUMEX. ACT is currently at the state where it can be shipped to other INTERLISP facilities. We have received a number of inquiries about the ACT system. ACT is a system in a continual state of development but we periodically freeze versions of ACT which we maintain and make available to the national AI community.

C. Highlights of Research Progress

Our ACTF system is a production system that operates in a semantic network data base. Our learning work has been focused on ways of increasing the power of production systems for performing various tasks. One class of learning mechanisms concern what we call knowledge compilation. This involves automatic mechanisms for creating productions that directly perform behavior that formerly required interpretative processing of knowledge in the semantic network. These compilation mechanisms also model the process by which human experts develop special purpose procedures to deal with the different types of problems that occur in their domain of expertise.